Course 27411 – Exercises 11/2 2008

PCA Exercise 1, Fisher *Iris* data

The Fisher Iris data-set is classical[1,2]. There are 150 objects, 50 *Iris setosa*, 50 *Iris versicolor* and 50 *Iris virginica*. The flowers of these 150 plants have been measured by a ruler. The variables are sepal length, sepal width, petal length and petal width (4 variables).

The original hypothesis was that *I. versicolor* was a hydrid of the two other species i.e. *I. setosa* x *virginica*. *I. setosa* is a diploid species; *I. virginica* is a tetrapoid and *I. versicolor* is a hexaploid.

1. Start The Unscrambler and open the help function. Lookup PCA and read about the different options and plots provided by the software.
   In general: Use the help function before asking! ☺ To get help on the different plots mark the plot and press F1.
2. Import the excel file fisherout.xls to UNSCRAMBLER and save the UNSCRAMBLER file
3. Usually you should examine the raw data first, but in this exercise try to perform a PCA, with "leverage correction" and with centering. Look at the four standard plots.
   a. How many principal components would you need and what does the PC1 describe?
   b. How many % of the variation is described by the first two PCs?
   c. Do you see anything "wrong" in the score-plot?
   d. If so what is wrong and what should be done?
   e. Try to correct it!
4. Does it help to standardize the data matrix and perform a new PCA (try it!).
5. How many PCs are needed to explain 70, 75% or 90% of the variance?
6. What does PC1, PC2 and PC3 show?
7. Are any variables more discriminative than others? Are any variables dispensable?
8. Can you see the presupposed classes? Any class overlap?

[1] Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* 7: 179–188.
[2] Edgar Anderson (1935). "The irises of the Gaspe Peninsula". *Bulletin of the American Iris Society* 59: 2–5.

PCA exercise 2 (IMPORTANT – to be presented by Team 1 Monday 18/2 2008)

The second dataset is called vinout. It is 178 objects (Italian wines), the first 59 are Barolo wines, the next 71 are Grignolino wines and the last 48 are Barbera wines. These wines have been characterized by 13 variables:

1) Alcohol (%),
2) malic acid,
3) ash,
4) alkalinity of ash,
5) magnesium,
6) total phenols,
7) flavanoids,
8) nonflavanoid phenols,
9) proanthocyanins,
10) colour intensity,
11) colour hue,
12) OD280/OD315 of diluted wines,
13) Praline

Questions:

1. Import the vinout.txt
   (flat ASCII data file, give number of objects and variables before final importing)
2. Look at the raw data. Any severe outliers?
3. Try PCA without standardization (centering, cross-validation)
   a. Do you see any groupings?
   b. Is there any overlap between classes?
   c. Are any variables especially important?
4. Try PCA with standardization.
   a. Interpret the four standard plots.
   b. Are any variables especially important and are any of them dispensable?
   c. How many PCs should be retained? (and saved in the model)?
5. Suppose that alcohol % and proanthocyanins were especially healthy, which wine would you recommend?
6. Use the UNSCRAMBLER jack-knife method to test for significance of the variables
   a. Can you explain what the method does?